Music Conversion from Synthetic Piano to Chinese Guzheng Using Image-based Deep Learning Technique

Ping-Hsuan Chen, Kai-Wen Liang, and Pao-Chi Chang
National Central University, Taoyuan City, Taiwan

Abstract—In this paper, we propose music conversion using deep learning method to improve generation of Chinese Guzheng music. Based on human perceptual evaluation, the average score is up to 4.3 which is the similarity of Chinese Guzheng music generated by our method and real music from their impression. The generated Chinese Guzheng music preserves the feature of the real play. Moreover, this method provides a simple way which can be used by whom never learn any instrument. This is a brand new research with good results.

I. Introduction

Chinese Guzheng music is popular from ancient Qin dynasty [1]. However, from the study of modern Chinese history, Chinese people continued to pursue westernization. The Chinese Guzheng also faces the drastic changes of the external environment such as the needs of modern national orchestras. Due to changes, the Guzheng is reformed in the direction of expanding the sound range, increasing the volume [2]. Although new design for the Chinese Guzheng instrument has been proposed, it is too hard for players to catch the music transition because of long time producing and learning new skills.

To improve this issue, we propose music conversion with CycleGAN [3] which is a deep learning technique that involves the automatic training of image-to-image translation models without paired examples. Researchers have applied CycleGAN for music conversion, for example, in 2018, Gino Brunner et al. proposed MIDI-VAE [4] and Sicong Huang et al. proposed TimbreTron [5]. In this paper, it is the first time to adopt CycleGAN to generate western instrument music from Chinese Guzheng. Chinese Guzheng music is very different from eastern music. One of the prominent characteristics in playing Chinese musical instruments is the emphasis on "yun" [6]. Melodic complexity of western music is often greater, with the extensive use of bent notes and microtones, and polyrhythms and cross-rhythms play a much more significant role. The complexity makes automatic training difficult.

The contribution of this paper is not only preserving the feature of real playing but also propose a simple way for generating Chinese Guzheng music. We invite 20 people who don't have experience of playing Chinese Guzheng and 20 people who have experience of playing Chinese Guzheng to do the evaluation. The average score for similarity of Chinese Guzheng music generated by our method and real music is up to 4.3 based on a five-scale evaluation. Furthermore, this conversion can be done by whom have never learned any instrument but have music score. In this paper, we successfully improve the generation for new trend Chinese Guzheng music by music conversion.

II. Proposed Method

A. Preprocess

In this paper, the training data include two domains, one is synthetic piano and the other one is Chinese Guzheng. Synthetic piano audio is generated from Musescore which is a software for score writer [7]. Chinese Guzheng audio is collected from real playing. All collected songs are split into 5 seconds per file then sent to Photosounder to generate the BMP file of spectrogram. Photosounder [8] is an outline of phase vocoding which is a widely used type of granular synthesis [9]. The user can form audio-visual relationships through interacting with and observing perceptual similarities between sound and image at the micro and macro level. These relationships provide the user with a unified parameter for audio-visual manipulation. In this work, Photosounder is a preprocess and post-process tool to convert among audio signals and spectrogram images.

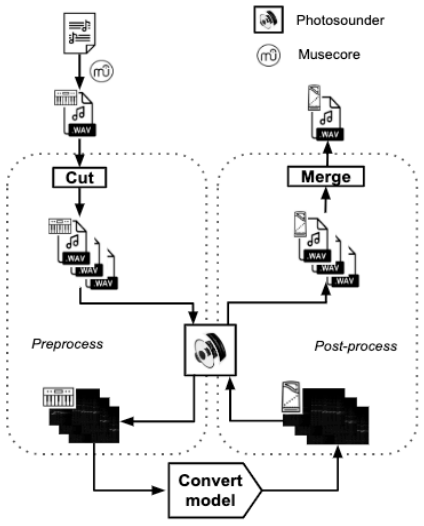

Fig.1 Preprocessing and Post-Processing

B. Architecture

In this work, we apply CycleGAN to generate Chinese Guzheng music. To make sure the quality of generating, this architecture includes generators and discriminator (Fig.2). These two kinds of model keep trying to beat each other until the generator creates nearly indistinguishable data from the real dataset. One generator called $G_{SynthPiano\_Guzheng}$, it contributes to transfer images from synthetic piano domain to Guzheng domain. The other generator is called $F_{Guzheng\_SynthPiano}$, it contributes to convert images back to synthetic piano domain from Guzheng domain. The generator is an encoder-decoder model architecture. The system includes two stride-2 convolutions, nine residual blocks for the training images

978-1-7281-7399-3/20/$31.00 ©2020 IEEE

higher than 256 × 256 resolution, and two ½ -stride convolutions with fractionally stride. The residual blocks are done to ensure the properties of previous layers are retained for later layers. For the discriminator, $D_{Guzheng}$, 70 × 70 PatchGAN is used to classify the patches of image is real or fake. So far, the models are sufficient for generating plausible images in the target domain (Eq.1). The cycle consistency loss calculates the difference between the image input to $G_{SynthPiano\_Guzheng}$ and the image output by $F_{Guzheng\_SynthPiano}$ and the generator models are updated accordingly to reduce the difference in the images (Eq.2). The full objective function is as Eq.3, the λ is suggested setting 10 and learning rate for all networks is 0.0002. Because of unsupervised manner using a collection of images from the source and target domain that do not need to be related in any way, this architecture is very suitable for Chinese Guzheng which is played by a variety of ways, such as alternating left and right hands, and multiple voices [10].

$$Loss_{adv}(G, D, data) = \frac{1}{m}\sum_{i=1}^{m}[1 - D(G(data))]^2 \text{--(1)}$$

$$Loss_{cyc}(G, F, data_x, data_y) = \frac{1}{m}\sum_{i=1}^{m}[F(G(data_x)) - data_x] + [G(F(data_y)) - data_y] \text{--(2)}$$

$$Loss_{full} = Loss_{adv} + \lambda Loss_{cyc} \text{--(3)}$$

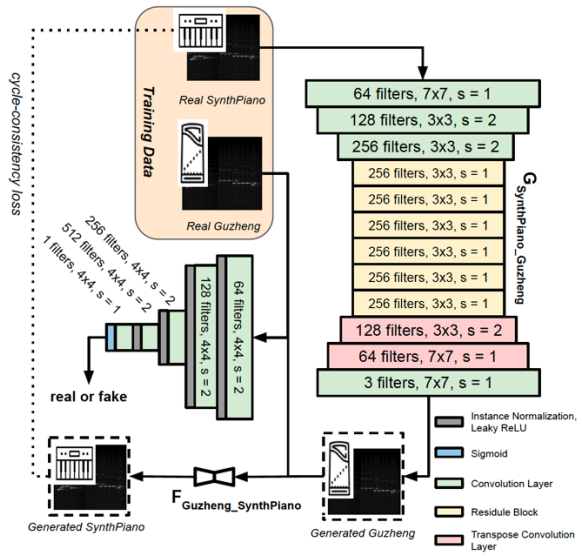

Fig.2 Training Architecture

III. Experiment

The collected songs are list in Table I. Every song is collected in 3 keys and tempo is 88 beats per minute with synthetic piano and Chinese Guzheng.

Chinese Name	English Name	Total Time(s)
雨夜花	*Flowers in the Rainy Night*	16.4
六月茉莉	*Jasmine in June*	21.8
望春風	*Spring Wind*	43.6
天黑黑	*Cloudy Day*	27.3
草螟弄公雞	*Grasshopper Playing Tricks on a Rooster*	27.3
滿山春色	*Springtime Hills*	38.2
桃花過渡	*Peach Blossom Takes the Ferry*	32.7

Table.1 Collected Songs

IV. Evaluation

We adopt human perception to investigate whether the method of this paper has the ability of preserving the musical content. The questionnaire is let experiment subject to listen the generated Guzheng music from our proposed method then give a score about the similarity with their impression of Chinese Guzheng music. The score is from 1 to 5 and the higher score means the higher similarity.

Score \ Subject	20 people ***have experience for playing Guzheng***	20 people ***no experience for playing Guzheng***	**Total experiment subject**
Average score	4.1	4.5	**4.3**

Table.2 Average score

V. Conclusion

In this paper, we propose music conversion to improve generation of new trend of Guzheng music. The evaluation shows that the method has the ability to generate a good result. Moreover, it is convenient everyone can use this method to generate Chinese Guzheng music even he cannot play instrument. While this model is just for generating Guzheng music, more flexible architecture should be explored as well. Also, the conversion between audio and image can be improve by deep learning method in the future.

VI. References

[1] J. R. Yeh & J. K. Hsu. (2015). Tradition, Evolution, and Regeneration: The Study of East Asian Contemporary Zheng Compositions from the Perspective of Acculturation. China Cultural University, Taipei, Taiwan. Retrieved from https://hdl.handle.net/11296/757vrs/

[2] H. Y. Tsai and H. Y. Huang. (2010). The Construction and Performing Techniques of the S-shaped 21-stringed Zheng and the Well-temperted New Aheng: A Comparative Study. Retrieved from https://hdl.handle.net/11296/92be59/

[3] J. Y. Zhu & T. Park & P. Isola & A. Efros. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2242-2251. 10.1109/ICCV.2017.244

[4] G. Brunner & A. Konrad & Y. Wang & R. Wattenhofer. (2018). MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer. arXiv :1809.07600

[5] Si. Huang & Q. Li & C. Anil & X. Bao & S. Oore & R. B. Grosse. (2018). TimbreTron: A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer. arXiv:1811.09620

[6] J. H. Shen & C. C. Liu. (2010). 古箏音樂的滑音自動識別. Retrieved from https://hdl.handle.net/11296/4583s3/

[7] Musescore, (n.d.) Retrieved from https://musescore.org/en/

[8] Photosounder. Retrieved from http://photosounder.com/

[9] J. S. Walker & G. W. Don. Mathematics and Music: Composition, Perception, and Performance, p.241-253. ISBN 1439867097, 9781439867099

[10] H. P. Wang (2013). 中国民族乐器简编. Beijing. ISBN 7516607800, 9787516607800